Uncertainty quantification for mean estimates

Author:
Noémi Takács

Supervisor:
Ambrus Tamás
SZTAKI, ELTE

Math Project ELTE, January 2024

Table of Contents

Introduction

Topic of the project: confidence region estimates
First semester: one-dimension, i.e. confidence intervals

- in scalar case, the general linear regression model:

  $$Y_t = X_t \cdot \vartheta^* + \varepsilon_t \quad (t = 1, \dots, n)$$

- constant in the noise: $X \equiv 1$
- assumptions on the noise term:
  - independence
  - symmetry
- confidence intervals:
  - inclusion
  - length

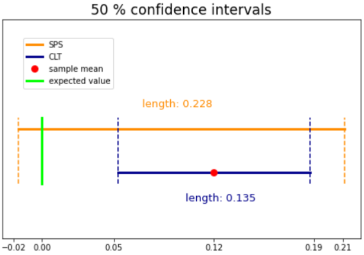

Figure: Example with a uniform(-1;1) sample, n = 30

The SPS method

Advantages:

- mild statistical assumptions
- distribution-free
- non-asymptotic confidence intervals

Main idea of the SPS:

- introduce sign-perturbed sums $\{S_i(\vartheta)\}$ and a reference sum $S_0(\vartheta) \doteq \sum_{t=1}^{n} \varepsilon_t(\vartheta)$
- construct a confidence interval based on the rank of $S_0(\vartheta)$

Theorem

Assuming the independence and the symmetry about zero of the noise term, the coverage probability of the SPS confidence interval is exactly p, where p is the user-chosen confidence level.

Simulations

Steps of the simulations:

- generate a sample
- construct confidence intervals (50 %), using SPS and a CLT based method
- repeat 10 000 times for $n = 10, 20, 30, \ldots, 100$

Measurements:

- inclusion rate of the true parameter
- average length of the intervals

Examined distributions:

- Standard normal
- Mixture of two normal: $P(X \in \mathcal{N}(m, 1)) = P(X \in \mathcal{N}(m, -1)) = 0.5$ for $m = 2, 10, 20$
- Student's t with df $= 2$
- Standard Cauchy
- Symmetrized Pareto with $\alpha = 2.5, 1.5, 0.5$

Results

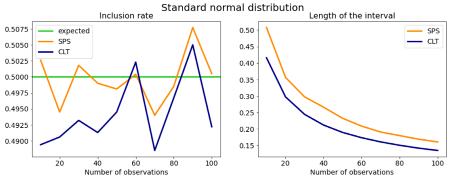

Figure: Simulation results for standard normal distribution

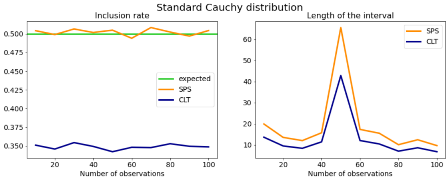

Figure: Simulation results for standard Cauchy distribution

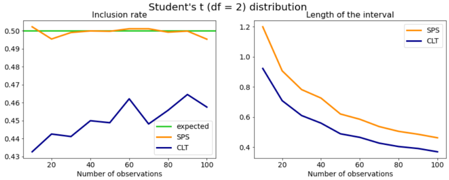

Figure: Simulation results for t distribution

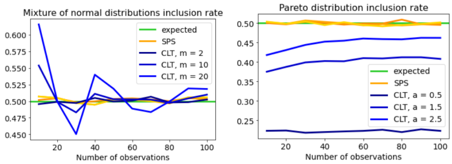

Figure: Simulation results for normal mixtures and Pareto distributions

Classification

Difference w.r.t. regression problems: the output is discrete.
In binary case let $Y_t \in \{0, 1\}$.
No explanatory variable $\Rightarrow Y_t \sim Ind(\theta^*) \Rightarrow n\bar{Y} \sim Bin(n, \theta^*)$

My simulation:

- $\theta^* = 0.8$
- confidence level = 50 %
- Methods based on
  - SPS
  - CLT
  - binomial distribution (expected to be the best)

Results

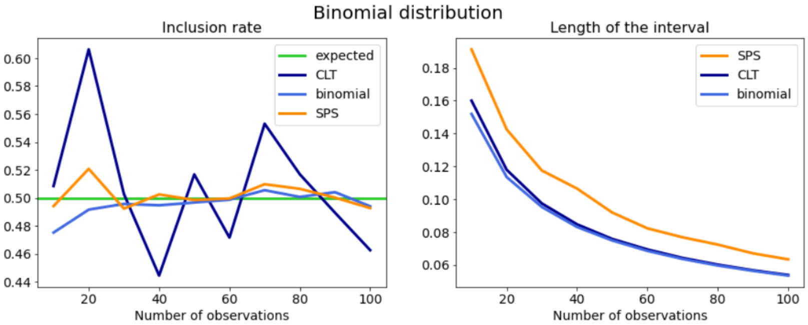

Figure: Simulation results for binomial distribution

References

- Csáji, B. Cs., Campi, M. C., Weyer, E. (2015). Sign-Perturbed Sums: A New System Identification Approach for Constructing Exact Non-Asymptotic Confidence Regions in Linear Regression Models. *IEEE Transactions on Signal Processing*, 63(1), 169–181.
- Szentpéteri, Sz., Csáji, B. Cs. (2023). Sample Complexity of the Sign-Perturbed Sums Identification Method: Scalar Case*. *IFAC-PapersOnLine*, 56(2), 10363-10370.

Thank you for your attention!